Inforex — a Collaborative System for Text Corpora Annotation and Analysis Goes Open

Michał Marcińczuk and Marcin Oleksy
G4.19 Research Group
Department of Computational Intelligence
Faculty of Computer Science and Management
Wrocław University of Science and Technology, Wrocław, Poland
{michal.marcinczuk,marcin.oleksy}@pwr.edu.pl

Abstract

In the paper we present the latest changes introduce to Inforex — a web-based system for qualitative and collaborative text corpora annotation and analysis. One of the most important news is the release of source codes. Now the system is available on the GitHub repository (https://github.com/CLARIN-PL/Inforex) as an open source project. The system can be easily setup and run in a Docker container what simplifies the installation process. The major improvements include: semi-automatic text annotation, multilingual text preprocessing using CLARIN-PL web services, morphological tagging of XML documents, improved editor for annotation attribute, batch annotation attribute editor, morphological disambiguation, extended word sense annotation. This paper contains a brief description of the mentioned improvements. We also present two use cases in which various Inforex features were used and tested in real-life projects.

1 Introduction

Development and evaluation of tools for various natural language processing task (named entity recognition, sentiment analysis, cyberbully detection and many other) require dedicated resources in a form of manually or semi-automatically annotated corpora. Corpus-based studies in the domain of Digital Humanities also require a support in the form of specialized tools and system. Both create a demand on development of tools and systems qualitative text corpora management, annotation, analysis and visualization.

Inforex is one of several web-based systems for text corpora annotation which is being developed as an open source project. The other well-known systems include, but are not limited to, WebAnno 3.0 (de Castilho et al., 2016), Brat (Stenetorp et al., 2012) and Anafora (Chen and Styler, 2013). Comparing to the other systems Inforex has some distinct features, including: support for untokenized and tokenized documents, support for both plain text and XML documents (XML tags are used to format the document layout) and integration with CLARIN-PL web services (utilizes on-demand morphological tagging).

In Section 2 we present the basic characteristic of the Inforex system. In Section the 3 we present the recents improvements and new features implemented in the system. In the Section 4 we present two projects in which the latest features were utilized.

2 Inforex Overview

Inforex is a web-based system for text corpora management, annotation and analysis. Since 2018 it is available as an open source project on the GitHub repository and is a part of the Polish CLARIN infrastructure[1] — it is integrated with the official repository for language resources in Polish CLARIN[2]. From the user perspective Inforex requires only a modern web browser to use the system.

Inforex offers several features for collaborative work on a single corpus, including concurrent access to data stored in the central database, role-based access to different modules, flag-based mechanism to track the process of various types of tasks. It support text cleanup, mention annotation, relations between annotations, morpholog-

[1] https://inforex.clarin-pl.eu
[2] https://clarin-pl.eu/dspace/

Proceedings of Recent Advances in Natural Language Processing, pages 711–719,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_083

ical tagging, annotation attributes, metadata and many others. A more comprehensive list of functions can be found in (Marcinczuk et al., 2017).

3 Recent Changes and Improvements

3.1 Open Source Project

After 10 years of development the project has been finally released as an open source project. The source codes are available under the LGPL license and can be obtained from https://github.com/CLARIN-PL/Inforex.

3.2 Easy Installation

The installation process was simplified by converting the system and all required components to run withing a set of Docker containers[3] defined in a Compose file. The Compose[4] file defines four containers: (1) **www** — web server running the Inforex application with background services (see Section 3.3), (2) **db** — MySQL database server, (3) **liquibase** — Liquibase database schema control and (4) **phpmyadmin** — web-based access to the database (for development and maintenance purposes).

A new installation of Inforex boils down to running the following two lines of code:

```
sudo apt-get install composer \
        docker docker-compose
./docker-dev-up.sh
```

3.3 Background Processes

Time consuming tasks, like corpus export or morphological tagging, are handled by processes running in the background. That's how we avoid the web server timeouts and handle task queuing. The background processes have been added to the Docker container with web server and they are automatically run on the container startup.

3.4 Multilingual Morphological Tagging

Inforex uses CLARIN-PL Web Service API[5] (Walkowiak, 2018) to facilitate the on-demand morphological document tagging. CLARIN-PL WS API provides access to morphological taggers for 11 languages. Seven of them are available from Inforex, i.e. Polish, English, German, Russian, Hebrew, Czech and Bulgarian (see Figure 1). Inforex automatically choose the language specific tagger based on the document language set in the metadata.

3.5 Extended Annotation Attribute Editor

We have extended the annotation attribute editor to handle dictionary-based attributes with a large number of possible values (see Figure 2). The improvements include the following:

- Filtering of the list of values,
- Feature to add a new element to the dictionary directly from the value picker level.
- Suggestions based on values assigned to other annotations. We have implemented two heuristic of generating the candidates with different levels of certainty, i.e.:
  - values for other annotations matched by the text form with the Soundex algorithm[6]. The list of candidates is sorted by their frequency,
  - attribute values matched by the annotation text (full or partial matching).

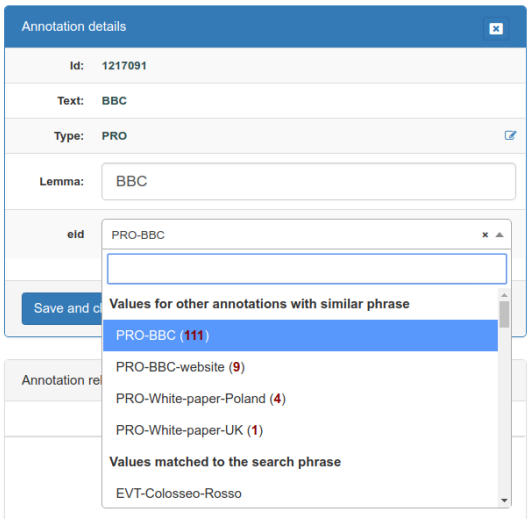

Figure 2: Extended annotation attribute editor

3.6 Batch Annotation Attribute Editor

Up to now the modification of annotation attributes was available only from the *Annotator* perspective using the annotation editor (see Figure 2). When an user had to modify an attribute for each annotation the only way was to go through all the annotations one by one. This process was time

[3] https://www.docker.com/
[4] https://docs.docker.com/compose/
[5] http://ws.clarin-pl.eu/tagerml.shtml
[6] https://www.archives.gov/research/census/soundex.html

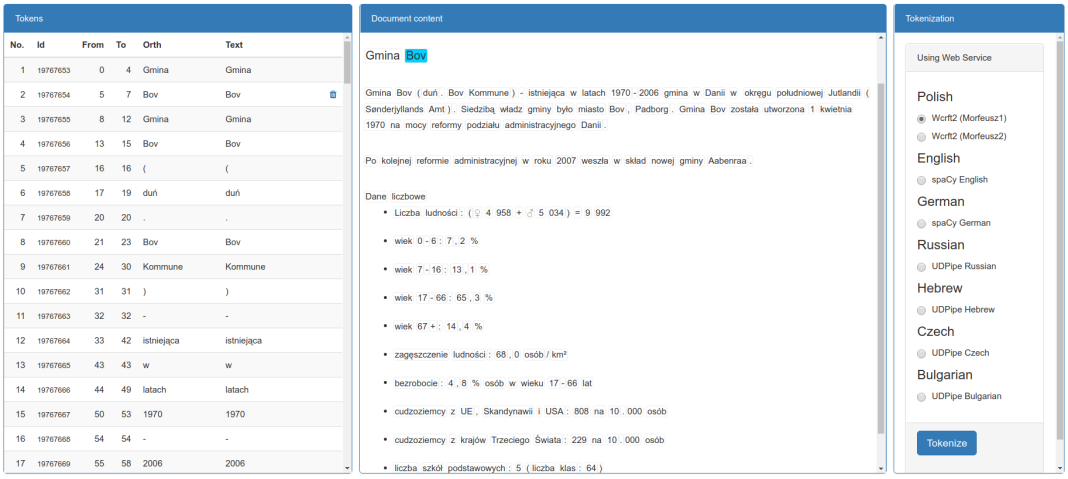

Figure 1: Document tokenization perspective

consuming and error-prone because it was easy to miss some annotations. To overcame these problems we have created a page for batch annotation attribute modification. The page consists of two main components, i.e. a document content with annotation preview and a table with annotations with their attribute (see Figure 3).

3.7 Document Auto Annotation

This feature was designed to reduce user effort in annotating repeatable phrases in and across documents. *Auto annotation* works by annotating in given documents all phrases that were already annotated in other documents. This feature works for both untokenized and tokenized documents, however we advise to use it on tokenized documents as the phrases are aligned with token boundaries and we avoid matching of incomplete words. After running *auto annotation* the new annotations are presented to the user for verification. User can decide whether given annotation is correct, incorrect or the annotation type needs a change (see Figure 4). The discarded annotation are stored in the system for future run of *auto annotation*. During the next use of *auto annotation* the new annotations which were previously discarded are ignored.

3.8 Lemma and Attribute Auto Fill

These features were designed to reduce user effort in setting annotation lemmas and attribute values. They both works in a similar way — for each annotation in the document the lemma or attribute value is set based on other annotations in the corpus. For lemma we collect annotations with the same text form and category. For attribute value we collect annotations with the same text form or lemma and category. In case of ambiguity, i.e. there are more than one possible value of lemma or attribute, the value remain empty and the user has to fill it manually. The lemma auto fill feature is available in the *Annotation lemmas* perspective and the attribute auto fill feature is available in the *Annotation attributes* perspective.

3.9 Tokenization of XML documents

Inforex allows to store documents in one of the two formats: plain text or XML. The XML format is used to encode document structure, like in the KPWr (Broda et al., 2012) and PCSN (Marcińczuk et al., 2011) corpora. During tagging the XML tags should be ignored and only the content should be processed. Thus, we made the tokenization process to be aware of the document format (see Figure 1). For XML format the document content is cleaned from XML tags, than the content is processed by the tagging service and at the end the tokenization is aligned with the original XML document.

3.10 Annotation Attribute Browser

The attribute value browser (see Figure 5) allows to browse corpus annotations by given attribute value. The page consists of three elements:

- View configuration — provides a set of filters, including: shared attribute, document language and subcorpus,
- Attribute values assigned to annotations — list of values their frequency,

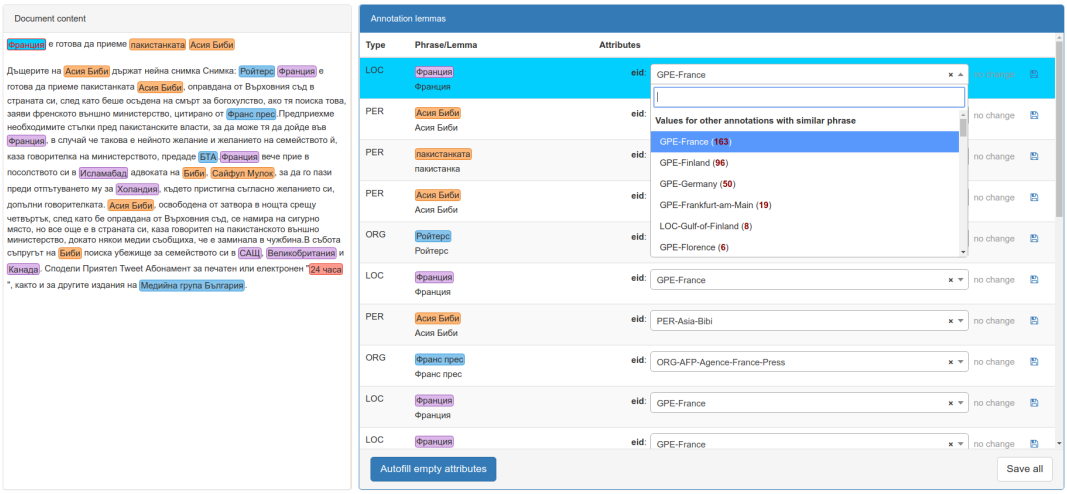

Figure 3: Batch annotation attribute editor

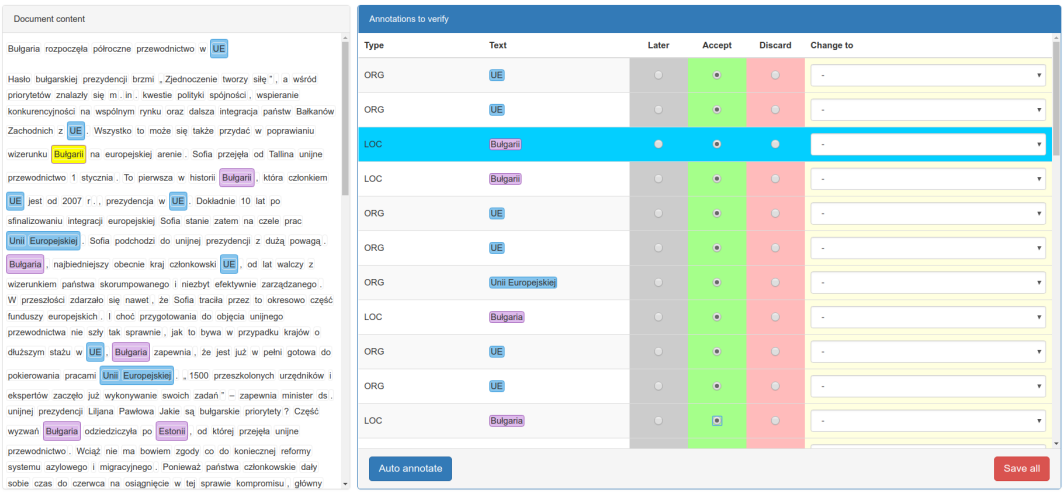

Figure 4: Auto annotation and candidate verification perspective

- Annotations with the selected value.

3.11 Export of Morphological Tagging and Annotations Agreements

For tokenized document Inforex can store up to three layers of morphological tags:

- *Tagger* — tags produced by a tool,
- *Agreement* — tags entered by a user in the agreement mode,
- *Final* — tags approved by the super user.

During export it is possible to define which layer of tags should be exported. It is possible to choose one of the following options (see Figure 6):

- *Final or tagger (if final not present)* — export the *final* tags. For tokens which does not have the *final* tag a *tagger* tag is taken.
- *Final* — export only the *final* tags. If there are tokens without final tags than the missing tags are reported as errors.
- *User (agreement)* — export tags created by selected user. For tokens which does not contain user agreement tags the tagger tags are taken.
- *Tagger* — export tagger tags.

3.12 Improved Support for Word Sense Annotation

The existing mechanism for word sense annotation was limited to a single set of words and their senses (Marcińczuk et al., 2012). We have removed the limitation and allow to define and use any number of sets of word senses in the WSD

Figure 5: Annotation attribute browser

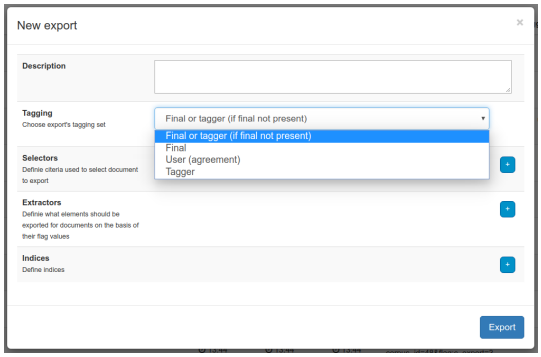

Figure 6: Export configuration dialog window

perspective. We also added the option to annotate the word senses in an agreement mode for further agreement (see Figure 7). Finally, we have imported all lexical units with their senses from Słowosieć 3.2 (Piasecki et al., 2016) as an annotation set to Inforex.

3.13 Morphological Agreement

The last feature is support for morphological disambiguation agreement. Inforex provides a page with morphological tag agreement across a given set of documents (see Figure 8). The agreement is presented in a numerical form for each documents in the set and after a specific document a list of disagreements is presented. This feature is complemented by a document perspective for comparing and choosing the final morphological tags (see Figure 9).

4 Case Studies

In this section we present two uses cases in which various features of Inforex were used in real-life projects.

4.1 BSNLP 2019 Shared Task

Inforex was used to create the training and testing datasets for the need of 2nd Edition of the Shared Task on Multilingual Named Entity Recognition for Slavic languages[7]. The task aims at recognizing mentions of named entities in news articles in Slavic languages, their lemmatization, and cross-language matching.

More than 10 people were involved in the annotation process for four languages, i.e. Polish, Czech, Russian and Bulgarian. There were 1–3 annotators per language. The annotation process consists of four main steps:

1. Selection of relevant documents — the document were automatically crawled and uploaded to Inforex, therefore some of them were duplicates or text not relevant to the subcorpus topic. The selected documents were marked with a flag *Valid content*.

2. Annotation of named entity mentions — the same set of five annotatation types was used to annotated all the selected documents. For Polish and Czech the annotators utilized the auto annotate feature described in Section 3.7 and we were able to evaluate the usability of

[7] http://bsnlp.cs.helsinki.fi/shared_task.html

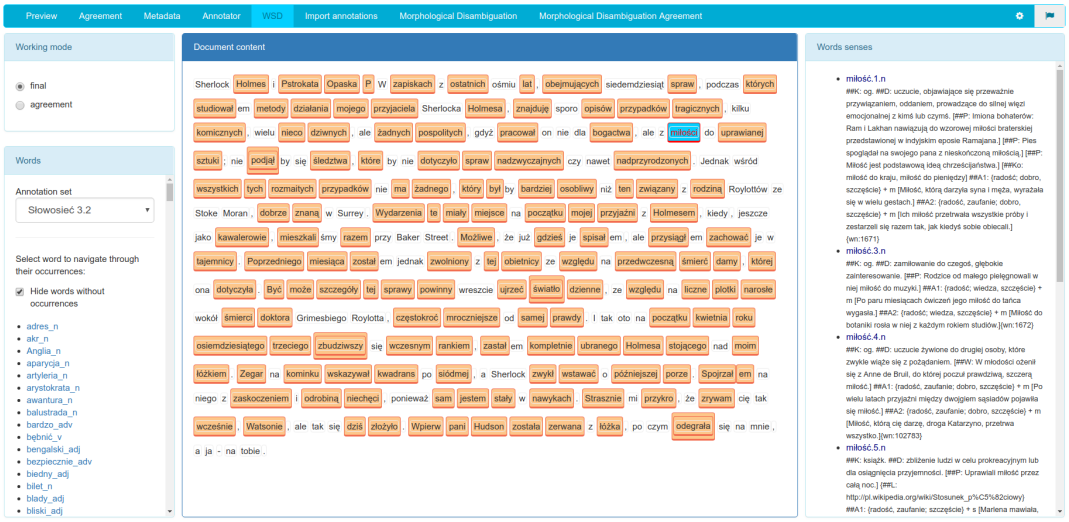

Figure 7: The extended perspective for word sense annotation

the auto annotation feature. Table 1 contains evaluation of the automatically recognized and added annotations for two languages and two subcorpora (each on a different topic). The auto annotation feature yielded very high precision of 97-99% with relatively high recall of 66-82%. This means that in case of Polish and Czech subcorpora 10k out of 14k annotations were added automatically. It was a significant facilitation of the work.

3. Assignment of annotation lemmas — to assign lemmas the annotators utilized batch lemma editor and lemma auto fill. For the correctly added annotations the lemmas were also automatically assigned.

4. Assignment of cross-lingual identifier — the goal was to assign the same identifier for each mention across all languages referring to the same real-world entity. There were more than 4k identifiers. The annotators utilized the auto fill features described in Section 3.8. The attribute browser was used to validate the entity mentions.

Language	Polish		Czech	
Subcorpus	A	B	A	B
Total	5139	2440	4183	2504
Final	5032	2386	4128	2502
Discarded	107	53	55	2
Add by user	1015	648	696	829
Precision [%]	97.4	97.0	98.4	99.9
Recall [%]	79.9	72.8	83.1	66.9

Table 1: Evaluation of the *auto annotation* feature on the BSNLP 2019 Shared Task dataset

4.2 Polish Translation of the NTU Multilingual Corpus

An ongoing project which goal is to provide Polish translation of the NTU Multilingual Corpus (Tan and Bond, 2011) which consists of two stories from the Sherlock Holmes Canon (The Adventure of the Speckled Band and The Adventure of the Dancing Men. The Adventure of the Speckled Band is the first one translated and prepared for manual annotation. The text was divided into 31 samples (txt files) of a similar size and imported directly into Inforex system. Then automatic morphological tagging was performed using WCRFT morpho-syntactic tagger for Polish (Radziszewski, 2013). The tagger provided morphological disambiguation on the basis of its context but also other possible forms for this particular word were listed. The result of the automatic annotation was then verified by two linguists independently. They were able to see morphological analysis of each token and the decision of the tagger (see Fig. 9). It could be accepted or discarded by the human annotator. It was also possible to add and assign an interpretation which was not identified by the tagger (e.g. in the case of unknown words). Inter-annotator agreement was calculated and its level was high enough (0,97) to perform further. Then, after completion the manual verification of morphological tagging by both linguists team coordinator proceeded with inconsistencies analysis. The decision was made for every token differently

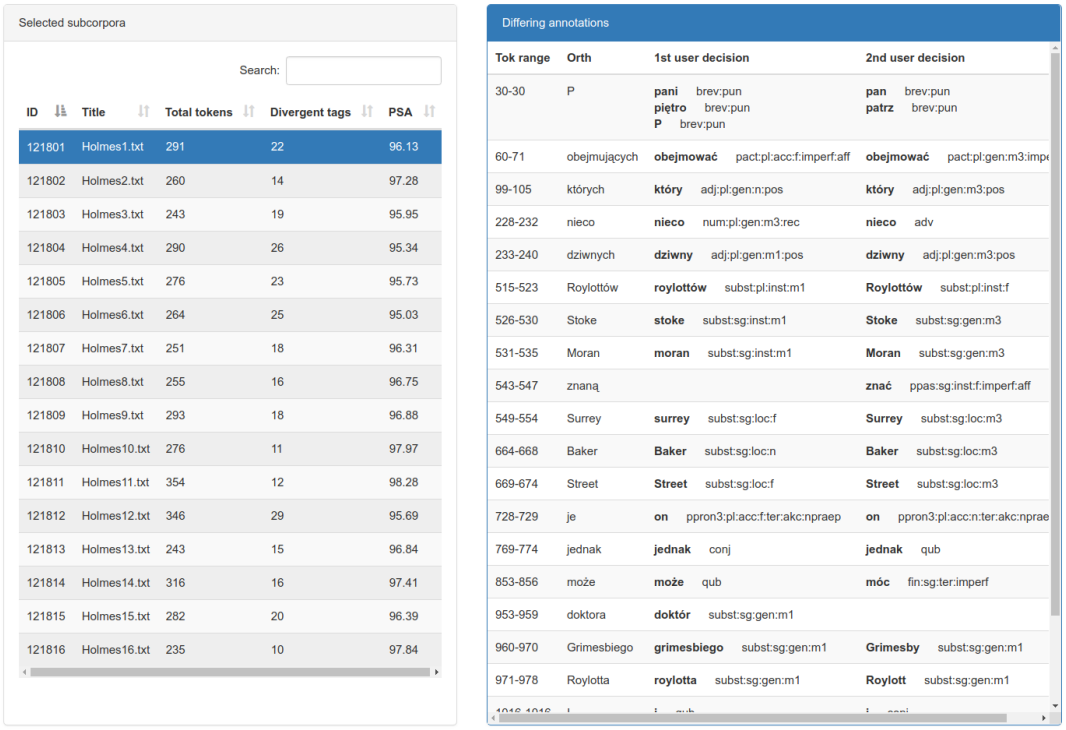

Figure 8: Summary of morphological disambiguation agreement

annotated. All tags verified by the team leader obtained the status of final annotations. They were added to the version published within CLARIN-PL infrastructure (Błaszczak et al., 2019).

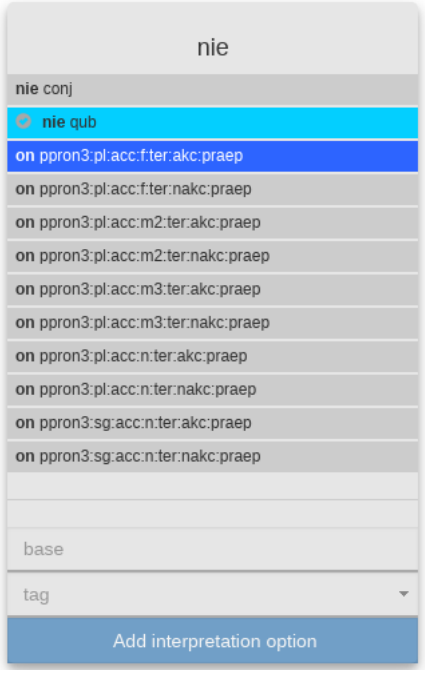

Figure 10: Morphological information provided for human annotators

After the Inforex functionality was developed and primarily used for creation of The Adventure of the Speckled Band corpus, it was successfully applied to prepare Corpus of the colloquial Polish language (Oleksy, 2019) in another project. This corpus has been designed to address the problem of morphological tagging of user-generated content (UGC) as part of the project "SentiCognitiveServices — next generation service for automating voice of customer and social media support based on artificial intelligence methods"[8]. The whole corpus (approximately 400000 tokens) is manually annotated with morphological information and furthermore the sample of 100 documents was prepared as a result of 2+1 annotation.

5 Summary

The last two years have been productive in the development of the Inforex system. Many new features and extensions were implemented during that time and the most important were presented in this paper. Majority of the features and improvements were dedicated by users. The most important news is the that Inforex has been finally released as an open source project.

Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the

[8]https://sentione.com/knowledge/eu-research-project

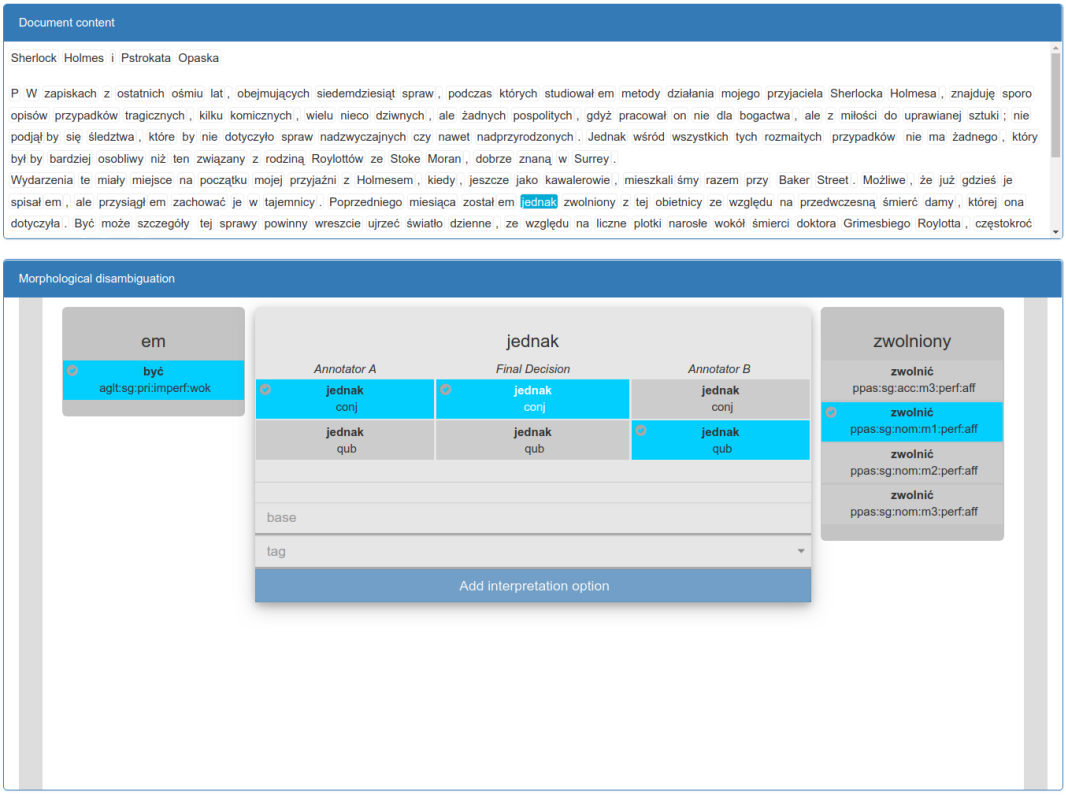

Figure 9: The perspective for morphological disambiguation agreement

Polish Ministry of Science and Higher Education.

References

Marta Błaszczak, Kacper Paszke, Ewa Rudnicka, Marcin Oleksy, Jan Wieczorek, Wioleta Kobylińska, Dominika Fikus, and Dagmara Kałkus. 2019. The adventure of the speckled band 1.0 (manually tagged). CLARIN-PL digital repository. http://hdl.handle.net/11321/667.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of LREC'12*. ELRA, Istanbul, Turkey.

Wei-Te Chen and Will Styler. 2013. Anafora: A web-based general purpose annotation tool. In *Proceedings of the 2013 NAACL HLT Demonstration Session*. Association for Computational Linguistics, pages 14–19. http://www.aclweb.org/anthology/N13-3004.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH) at COLING 2016*. pages 76–84. http://tubiblio.ulb.tu-darmstadt.de/97939/.

Michał Marcinczuk, Marcin Oleksy, and Jan Kocon. 2017. Inforex - a collaborative system for text corpora annotation and analysis. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*. INCOMA Ltd., pages 473–482. https://doi.org/10.26615/978-954-452-049-6_063.

Michał Marcińczuk, Monika Zaśko-Zielińska, and Maciej Piasecki. 2011. Structure annotation in the polish corpus of suicide notes. In Ivan Habernal and Václav Matoušek, editors, *Text, Speech and Dialogue*, Springer Berlin Heidelberg, volume 6836 of *Lecture Notes in Computer Science*, pages 419–426.

Michał Marcińczuk, Jan Kocoń, and Bartosz Broda. 2012. Inforex – a web-based tool for text corpus management and semantic annotation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Marcin Oleksy. 2019. Corpus of the colloquial polish language. CLARIN-PL digital repository. http://hdl.handle.net/11321/637.

Maciej Piasecki, Stan Szpakowicz, Marek Maziarz, and Ewa Rudnicka. 2016. PlWordNet 3.0 – Almost There. In Verginica Barbu Mititelu, Corina Forăscu, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the 8th Global Wordnet Conference, Bucharest, 27-30 January 2016*. Global Wordnet Association, pages 290–299.

Adam Radziszewski. 2013. A tiered crf tagger for polish. In *Intelligent Tools for Building a Scientific Information Platform*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*. Association for Computational Linguistics, Avignon, France.

Liling Tan and Francis Bond. 2011. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*. Institute of Digital Enhancement of Cognitive Processing, Waseda University, Singapore, pages 362–371. https://www.aclweb.org/anthology/Y11-1038.

Tomasz Walkowiak. 2018. Language processing modelling notation – orchestration of nlp microservices. In Wojciech Zamojski, Jacek Mazurkiewicz, Jarosław Sugier, Tomasz Walkowiak, and Janusz Kacprzyk, editors, *Advances in Dependability Engineering of Complex Systems*. Springer International Publishing, Cham, pages 464–473.